Approximate Distance Ranking-Based Validation for Spatial Contextual Classification: A Case Study of Election Data

Xiaorui Wei, Weiquan Zhao, and Yangping Li*

Dongguan University of Technology
tmhu@dgut.edu.cn

Abstract. Classification on spatial data is different from classical classification in that spatial context must be taken into account. In particular, the validation criterion functions should incorporate both classification accuracy and spatial accuracy. However, direct combination of the two accuracies is cumbersome, due to their different subjects and scales. To circumvent this difficulty, we develop a new criterion function that indirectly incorporates spatial accuracy into classification accuracy-based functions. Next, we formally introduce a set of ideal properties that an appropriate criterion function should satisfy, giving a more meaningful interpretation for the relative significance coefficient in the weighted scheme. Finally, we compare the proposed new criterion function with existing ones on a large data set for 1980 US presidential election.

Keywords: validation criterion function, distance ranking, spatial accuracy, spatial contextual classification.

1 Introduction

Spatial data mining [1,2] is an important component of data mining. In the case of spatial contextual classification, the class label of each site is not only determined by the local attributes, but is affected by its neighbors as well. For example, besides the house itself, the house price is also affected by the neighborhoods. Thus, we need to pay special attention to spatial context in both the model construction (training) phase and the model testing (validation) phase.

In this paper, we focus on a crucial component in the test phase, the validation criterion functions, which are used to evaluate the estimated class labeling against the true class labeling [3–5]. Appropriate validation criteria for geospatial data should capture both classification and spatial accuracies. However, traditional classification accuracies, such as classification rate, discard spatial information. One straightforward means of disambiguating the definition of a good multi-objective solution is to assign the accuracies different weights before combining them together. For instance, given

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Future Information Technology*,
Lecture Notes in Electrical Engineering 309,
DOI: 10.1007/978-3-642-55038-6_87, © Springer-Verlag Berlin Heidelberg 2014

classification accuracy C and spatial accuracy S, the weighted scheme gives $(1-\alpha)C+\alpha S$, where α is the relative significance coefficient, determined by the user according to his preference. However, because C is usually a classification rate-like measure with unit of percentage and S is a distance measure with unit of length, direct combination is cumbersome and can be even meaningless, due to their totally different subjects and scales. Besides, spatial accuracy computation is usually expensive, which involves nearest neighbor search and distance evaluation. To circumvent this difficulty, we propose a new criterion function that, instead of evaluating near neighbor distance values, approximates distance rankings using contiguity matrix. Its effectiveness is validated on a real-world database.

Table 1. Notations in the criterion functions

notation	Description
a_1	number of actual class 1 sites
a_0	number of actual class 0 sites, $n=a_1+a_0$
p_1	number of predicted class 1 sites
p_0	number of predicted class 0 sites, $n=p_1+p_0$
$\mathbf{a}_1$	$\mathbf{a}_1 \equiv \mathbf{z}$, actual vector for class 1,a_1=$\mathbf{a}_1^T\mathbf{a}_1$
$\mathbf{a}_0$	$\mathbf{a}_0 \equiv 1$-$\mathbf{z}$,actual vector for class 0,a_0= $\mathbf{a}_0^T\mathbf{a}_0$
$\mathbf{p}_1$	$\mathbf{p}_1 \equiv \hat{z}$,predicted vector for class 1,p_1=$\mathbf{p}_1^T\mathbf{p}_1$
$\mathbf{p}_0$	$\mathbf{p}_0 \equiv 1$- $\hat{z}$,predicted vector for class 0, p_0=$\mathbf{p}_0^T\mathbf{p}_0$

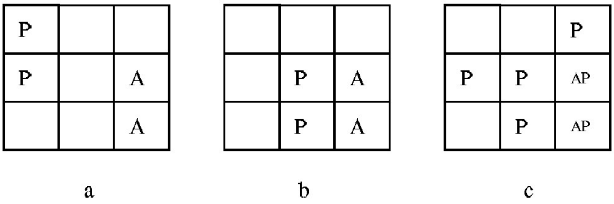

Fig. 1. CA vs ADNP. "A" denotes actual (class 1) site. "P" denotes predicted (class 1) site.

Overview. The rest of the paper is organized as follows. Section 2 introduces related work. Section 3 presents new criterion function. Section 4 discusses the ideal properties. Empirical results are reported in Section 5. Finally, Section 6 concludes this paper with some remarks.

2 Related Work

Geospatial classification problem can be informally described as follows. Given a spatial framework of n sites, where each site s_i has a class label $z_i \in \{0,1\}$ (we focus on binary classification in this paper) and a vector of explanatory attributes x_i, we need to construct a classification model to predict the class label. The class label of each site is not only determined by its own attributes, but also affected by its neighbors. The neighbor relationship *N* is often encoded in a contiguity matrix *W* for which W(*i*,*j*))> 0 if sites s_i and s_j are neighbors, W(*i*,*j*)=0 otherwise. In this paper, we focus on how to evaluate the similarity between the estimated binary class labeling $\hat{z} \equiv [\hat{z}_1,...,\hat{z}_n]$ and the true binary class labeling $z \equiv [z_1,...,z_n]$ .To ease the following discussion, we first introduce the relevant notations in Table1, where **1** is a *n*-D vector of 1's.

In classification, the most popular criterion is probably the classification accuracy, *CA*. As defined in Eq. (1), it simply computes the fraction of data that is correctly classified.

$$CA \equiv \frac{p_1^T a_1 + p_0^T a_0}{n} \tag{1}$$

For location prediction, however, spatial accuracy has to be considered and *CA* alone is not enough. See Fig. 1 for an example, where *CA* cannot distinguish the two predicted class labelings in Fig. 1a and b. Suppose we are interested in the sites of class 1, e.g., the location of gold mine. Domain scientist may prefer Fig.1b where predicted 1 locations are near actual 1 locations. In the case of gold mine, it is a reasonable expectation that even if we cannot accurately predict the real locations of gold mine, our predicted ones are not far away from them.

Along this line, as defined in Eq. (2), Reference [6] proposed a new measure, Average Distance to Nearest Predicted (class 1) location from actual (class 1) location (*ANDP*)

$$ADNP \equiv \sum_{s_i \in S_1} \frac{1}{a_1} d(s_i, NP(s_i)) \tag{2}$$

where S1 denotes the set of 1 sites, $\{s_i : a_1[i] = 1\}$, $NP(s_i)$ denotes the nearest predicted 1 site from s_i and $d(s_i, NP(s_i))$ denotes the distance between them. However, sometimes *ADNP* alone can lead to very low *CA* (see Fig. 1c) and always encourages more predicted 1 sites. In the extreme case where all sites are predicted 1, *ADNP* =0, a perfect but useless prediction.

3 Approximate Distance Ranking Function

Hence, for geospatial classification, a good criterion function should combine both classification accuracy and spatial accuracy *ADNP*. Reference[7] utilized the weighted scheme to define a new function *M*0 as

$$M0 \equiv (1-\alpha)CA + \alpha ADNP' \tag{3}$$

where α is a relative significance coefficient, *ADNP'* = exp(–*ADNP*) is normalized to [0,1].

However, while the normalization step is necessary, it is hard to select the appropriate normalization functions [8]. For instance, it is difficult to justify the use of exp(x) over $\frac{x-\min}{\max-\min}$, which may be domain-dependent or even dataset dependent. Besides, the computation of *ADNP* is not trivial, which involves nearest neighbor search and distance evaluation. In the worse case, both time and space complexities are is $O(n^2)$.

To trade accuracy for efficiency, we develop an approximate distance ranking function, which does not require finding nearest neighbors or computing distance. The motivation is as follows. What we care most is the relative ranking of estimated class labelings evaluated with *ADNP* , rather than the absolute distance values. We observe that the ranking can be approximated by utilizing the contiguity matrix *W* containing the neighborhood information for each site. The accuracy is not sacrificed much if sites are evenly distributed and there are predicted 1 sites in the neighborhood of every misclassified 1 site. Furthermore, *W* is a sparse matrix for which we only store those non-zero elements and thus a lot of storage space is saved.

For ease of expression, let $\mathbf{q} \equiv \mathbf{a}_1 \wedge \mathbf{p}_0$, where $\mathbf{a}_1$ and $\mathbf{p}_0$ are treated as boolean vector and $\wedge$ is bit-AND operator. q[*i*]=1 iff s_i is a 1 site but predicted 0.

$$M1 \equiv (1-\alpha)CA + \alpha(1-\frac{\mathbf{q}^T\mathbf{q}}{a_1}) \tag{4}$$

Function *M*1 is defined in Eq. (4). Instead of computing the average distance for those sites that are actually 1 but predicted 0, we just count those sites, in the hope that a smaller number of such sites would lead to a smaller distance.

4 Ideal Properties

To compare the various criterion functions for geospatial classification validation, we propose a series of ideal properties that a criterion functions must satisfy. In Eq. (3), the user-defined significance coefficient α is still confusing. In the ideal properties, we will give α a more meaningful interpretation.

In the proposed ideal properties, α is determined by the user and is typically small, e.g., 0.05. It provides a meaningful threshold, based on which we decide which measure, *CA* or *ADNP*, should be emphasized. If the absolute difference in *CA* is greater than α, the hypothesis with larger *CA* is favored, because a much larger *CA* often means a larger *ADNP*. If the absolute difference in *CA* is less than α, we ignore the slight difference in *CA* and favor the hypothesis with greater *ADNP*, i.e., spatial accuracy matters. Let $\mathbf{y}$, $\hat{\mathbf{y}}_1$, $\hat{\mathbf{y}}_2$ denote the actual class labeling and two predicted class labelings, respectively. Let $CA_i \equiv CA(\mathbf{y},\hat{\mathbf{y}}_i)$, $ADNP_i \equiv ADNP(\mathbf{y},\hat{\mathbf{y}}_1)$, where *i*=1,2. In detail, an ideal criterion function *M* combining *CA* and *ADNP* should satisfy the following properties.

1. If $|CA_1 - CA_2| > \alpha$, the estimation with larger *CA* is favored.
2. If $|CA_1 - CA_2| \le \alpha$, the estimation with larger *ADNP* is favored.

5 Experimental Evaluation

In this section, we first introduce the experimental setting. Then we use a large real-world election dataset to compare the criterion functions.

Table 2. Comparison of criterion functions

	Comparison with *M*0	Comparison with ideal properties	
criterion	*M*1	*M*0	*M*1
Rand	0.9958	0.6824	0.6865
Pearson	0.9913	0.3606	0.3637

5.1 Experimental Setting

As for the classification model, we select Spatial Autoregression Model (SAR) [9]. With the output from SAR, $\hat{\mathbf{y}} = P(Z_i = 1)$, we try 101 cutoff probability (threshold) values of θ sampled evenly at a constant interval 0.01 in [0,1], and transform $\hat{\mathbf{y}}$ to class labeling as $\hat{z}[i] = 1 \leftrightarrow \hat{y}[i] > \theta$. To emphasize the impact of spatial accuracy *ADNP* , we set α =0.2 for all criterion functions containing α as the coefficient. As for the similarity measure to compare the validation criterion functions, we use the pairwise ranking. In detail, given the true class labeling and a set of estimated class labelings, all possible pairs of estimates are taken and it is determined, whether they are treated in the same manner by two criterion functions. With these statistics, we can compute Rand index and Pearson coefficient. As for the dataset, we select a large real-world dataset for 1980 US presidential election results covering 3107 counties [9]. In the original regression problem, income, home ownership and population with college degrees are used to predict voting rate. We transform the original continuous target variable y to binary as z =1 if y> avg(y), z =0 otherwise.

5.2 Empirical Results

As we discussed previously, what we care most is the capability of *M*1 to approximate the pairwise ranking of *M*0, rather than the raw distance values or normalized values of *ADNP*. The left part of Table 2 reports their performance in terms of Rand index and Pearson coefficient. Apparently, both measures indicate a high resemblance.

The right part of Table 2 also gives the comparison with ideal properties. Although *M*0 employs the weighted scheme to explicitly incorporate *ADNP* and thus appears to satisfy the ideal properties (also specified in *ADNP*) best, it is not the best in either Rand index or Pearson coefficient. On the contrary, *M*1 becomes the best in this case, which justifies the simplifying assumptions we made to develop these approximate functions. Again, this confirms the usefulness and advantage of our proposed criterion functions to approximate ranking in *ADNP* .

6 Concluding Remarks

In geospatial classification validation, both classification accuracy and spatial accuracy must be taken into account. However, conventional criterion functions use the weighted scheme to directly combine the two accuracies, which are of different subjects and scales. In this paper, by leveraging the contiguity matrix that encodes the neighborhood relationship, we proposed an approximate distance ranking function that approximates the pairwise ranking of *ADNP*. It not only circumvents the difficulty in the direct weighted scheme, but also leads to smaller time and space complexity. Finally, its effectiveness was demonstrated on a large election dataset.

For the future work, we plan to investigate the relationship between *CA* and *ADNP* .We will also develop training procedures that specialize in optimizing the proposed criterion functions.

Acknowledgments. This work is supported by NSFC (No. 61100136) and SPIIERG (No. 2012B091100295).

References

1. Shekhar, S., Chawla, S.: Spatial Databases: A Tour. Prentice-Hall (2002)
2. Chun, Y., Griffith, D.: Spatial Statistics and Geostatistics: Theory and Applications for Geographic Information Science and Technology. SAGE (2013)
3. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, pp. 1137–1143 (1995)
4. Kim, J.: Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. Computational Statistics & Data Analysis 53(11), 3735–3745 (2009)
5. Schiavo, R.A., Hand, D.J.: Ten more years of error rate research. International Statistical Review 68(3), 295–310 (2000)
6. Shekhar, S., Schrater, P., Raju, W.R., Wu, W., Chawla, S.: Spatial contextual classification and prediction models for mining geospatial data. IEEE Transactions on Multimedia 4(2), 174–188 (2002)
7. Chawla, S., Shekhar, S., Wu, W.: Predicting locations using map similarity (PLUMS): A framework for spatial data mining. In: Proceedings of the International Workshop on Multimedia Data Mining, pp. 14–24 (2000)
8. Jain, A., Nandakumar, K., Ross, A.: Score normalization in multimodal biometric systems. Pattern Recognition 38, 2270–2285 (2005)
9. LeSage, J.P.: Bayesian estimation of spatial autoregressive models. International Regional Science Review 20, 113–129 (1997)